Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Importance Sampling with Unequal Support

Philip S. Thomas, Emma Brunskill
{philipt,ebrun}@cs.cmu.edu
Carnegie Mellon University

Abstract

Importance sampling is often used in machine learning when training and testing data come from different distributions. In this paper we propose a new variant of importance sampling that can reduce the variance of importance sampling-based estimates by orders of magnitude when the supports of the training and testing distributions differ. After motivating and presenting our new importance sampling estimator, we provide a detailed theoretical analysis that characterizes both its bias and variance relative to the ordinary importance sampling estimator (in various settings, which include cases where ordinary importance sampling is biased, while our new estimator is not, and *vice versa*). We conclude with an example of how our new importance sampling estimator can be used to improve estimates of how well a new treatment policy for diabetes will work for an individual, using only data from when the individual used a previous treatment policy.

Introduction

A key challenge in artificial intelligence is to estimate the expectation of a random variable. Instances of this problem arise in areas ranging from planning and decision making (e.g., estimating the expected sum of rewards produced by a policy for decision making under uncertainty) to probabilistic inference. Although the estimation of an expected value is straightforward if we can generate many *independent and identically distributed* (i.i.d.) samples from the relevant probability distribution (which we refer to as the *target distribution*), we may not have generative access to the target distribution. Instead, we might only have data from a different distribution that we call the *sampling distribution*.

For example, in off-policy evaluation for reinforcement learning, the goal is to estimate the expected sum of rewards that a decision policy will produce, given only data gathered using some other policy. Similarly, in supervised learning, we may wish to predict the performance of a regressor or classifier if it were to be applied to data that comes from a distribution that differs from the distribution of the available data (e.g., we might predict the accuracy of a classifier for hand-written letters given that observed letter frequencies come from English, using a corpus of labeled letters collected from German documents).

More precisely, we consider the problem of estimating $\theta := \mathbf{E}[h(X)]$, where h is a real-valued function and the expectation is over the random variable X, which is a sample from the target distribution. As input we assume access to n i.i.d. samples from a sampling distribution that is different from the target distribution. A classical approach to this problem is to use *importance sampling* (IS), which reweights the observed samples to account for the difference between the target and sampling distributions (Kahn 1955). Importance sampling produces an unbiased but often high-variance estimate of θ.

We introduce *importance sampling with unequal support* (US)—a simple new importance sampling estimator that can drastically reduce the variance of importance sampling when the supports of the sampling and target distributions differ. This setting with unequal support can occur, for example, in our earlier example where German documents might include symbols like ß, that the classifier will not encounter. US essentially performs importance sampling only on the data that falls within the support of the target distribution, and then scales this estimate by a constant that reflects the relative support of the target and sampling distributions.

US typically has lower variance than ordinary importance sampling (sometimes by orders of magnitude), and is unbiased in the important setting where at least one sample falls within the support of the target distribution. If no samples do, then none of the available data could have been generated by the target distribution, and so it is unclear what would make for a reasonable estimate. Furthermore, the conditionally unbiased nature of US is sufficient to allow for its use with concentration inequalities like Hoeffding's inequality to construct confidence bounds on θ. By contrast, *weighted importance sampling* (Rubinstein 1981) is another variant of importance sampling that can reduce variance, but which introduces bias that makes it incompatible with Hoeffding's inequality.

Problem Setting and Importance Sampling

Let f and g be *probability density functions* (PDFs) for two distributions that we call the *target distribution* and *sampling distribution*, respectively. Let $h : \mathbb{R} \to \mathbb{R}$ be called the *evaluation function*. Let $\theta := \mathbf{E}_f[h(X)]$, where $\mathbf{E}_f$ denotes the expected value given that f is the PDF of the random variable(s) in the expectation (in this case, just X). Let

$F := \{x \in \mathbb{R} : f(x) \neq 0\}$, $G := \{x \in \mathbb{R} : g(x) \neq 0\}$, and $H := \{x \in \mathbb{R} : h(x) \neq 0\}$ be the supports of the target and sampling distributions, and the evaluation function, respectively. In this paper we will discuss techniques for estimating θ given $n \in \mathbb{N}_{>0}$ i.i.d. samples, $\mathbf{X}_n := \{X_1, \ldots, X_n\}$, from the sampling distribution, and we focus on the setting where $F \cap H \subset G$—where the joint support of F and H is a *strict* subset of the support of G.

The importance sampling estimator,

$$\mathrm{IS}(\mathbf{X}_n) := t + \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)}(h(X_i) - t), \qquad (1)$$

is a widely used estimator of θ, where $t = 0$ (we consider non-zero values of t later). If $F \cap H \subseteq G$, then $\mathrm{IS}(\mathbf{X}_n)$ is a consistent and unbiased estimator of θ. That is, $\mathrm{IS}(\mathbf{X}_n) \xrightarrow{\text{a.s.}} \theta$ and $\mathbf{E}_g[\mathrm{IS}(\mathbf{X}_n)] = \theta$ (we review this latter result in Property 1 in the supplemental document).

A *control variate* is a constant, $t \in \mathbb{R}$, that is subtracted from each $h(X_i)$ and then added back to the final estimate, as in (1) (Hammersley 1960; Hammersley and Handscomb 1964). Although control variates, $t(X_i)$, that depend on the sample, X_i, can be beneficial, for our later purposes we only consider constant control variates. Intuitively, including a constant control variate equates to estimating $\theta' := \mathbf{E}_f[h'(X)]$ using importance sampling without a control variate, where $h'(x) = h(x) - t$, and then adding t to the resulting estimate to get an estimate of θ.

Later we show that the variance of importance sampling increases with θ^2, and so applying importance sampling to h results in higher variance than applying importance sampling to h' with $t \approx \theta$, since then $\theta' \approx 0$. That is, by inducing a kind of normalization, a control variate can reduce the variance of estimates without introducing bias—a property that has made the inclusion of control variates a popular topic in some recent works using importance sampling (Dudík, Langford, and Li 2011; Jiang and Li 2016; Thomas and Brunskill 2016). Although later we discuss control variates more, for simplicity our derivations focus on importance sampling estimators without control variates. There are also other extensions of the importance sampling estimator that can reduce variance—notably the weighted importance sampling estimator, which we compare to later, and which can provide large reductions of variance and mean squared error, but which introduces bias.

An Illustrative Example

In this section we present an example that highlights the peculiar behavior of the IS estimator when $F \cap H \neq G$. The illustrative example, depicted in Figure 1, is defined as follows. Let $g(x) = 0.5$ if $x \in [0, 2]$ and $g(x) = 0$ otherwise, and let $f(x) = 1$ if $x \in [0, 1]$ and $f(x) = 0$ otherwise. So, $F = [0, 1]$ and $G = [0, 2]$. Let $h(x) = 1$ if $x \in [0, 1]$ and $h(x) = 0$ otherwise, so that $H = [0, 1]$. Notice that $\theta = 1$.

Since the sampling and target distributions are both uniform, an obvious estimator of θ (if f and g are known but h is not) would be the average of the points that fall within F. Let $(\#X_i \in F)$ denote the number of samples in $\mathbf{X}_n$ that are in F. Formally, the obvious estimator is

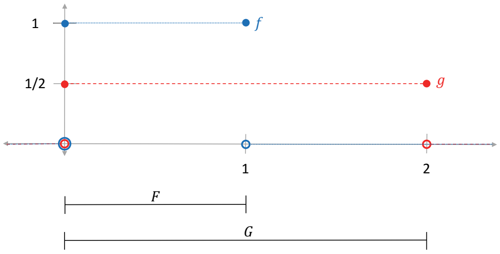

Figure 1: Depiction of the illustrative example. The evaluation function is not shown because $h = f$ and $H = F$.

$$\hat{\theta} := \frac{1}{(\#X_i \in F)} \sum_{i=1}^{n} \mathbf{1}_F(X_i)h(X_i),$$

where $\mathbf{1}_A(x) = 1$ if $x \in \mathcal{A}$ and $\mathbf{1}_A(x) = 0$ otherwise. Given our knowledge of h, it is straightforward to show that this estimator is equal to 1 if $(\#X_i \in F) > 0$ and is undefined otherwise—it is exactly correct (has zero bias and variance) as long as at least one sample falls within F. If no samples fall within F, then we have only observed data that will never occur under the target distribution, and so we have no useful information about θ. In this case, we might define our obvious estimator to return an arbitrary value, e.g., zero.

Perhaps surprisingly, the importance sampling estimator does not degenerate to this obvious estimator:

$$\mathrm{IS}(\mathbf{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_F(X_i) 2h(X_i) = \frac{2(\#X_i \in F)}{n}.$$

Since $\mathbf{E}_g[(\#X_i \in F)/n] = 1/2$, this estimate is correct in expectation, but does not have zero variance given that at least one sample falls within F. If more than $1/2$ of the samples fall within F, this estimate will be an over-estimate of θ, and if fewer than $1/2$ of the samples fall within F, this estimate will be an under-estimate. Although correct on average, the importance sampling estimator has unnecessary additional variance relative to the obvious estimator.

Importance Sampling with Unequal Support

We propose a new importance sampling estimator, *importance sampling with unequal support* (ISUS, or US for brevity), that *does* degenerate to the obvious estimator for our illustrative example. Intuitively, US prunes from $\mathbf{X}_n$ the samples that are outside F (or more generally, outside some set C, that we define later) to construct a new data set, $\mathbf{X}'_n$, that has fewer samples. This new data set can be viewed as $(\#X_i \in F)$ i.i.d. samples from a different sampling distribution—a distribution with PDF g', which is simply g, but truncated to only have support on F and re-normalized to integrate to one. US then applies ordinary importance sampling to this new data set.

For generality, we allow US to prune from $\mathbf{X}_n$ all of the points that are not in a set, C, which can be defined many different ways, including $C \coloneqq F$ (as in our previous example). Our only requirement is that $F \cap H \subseteq C \subseteq G$. In order to compute US, we must compute a value,

$$c \coloneqq \int_C g(x)\,\mathrm{d}x,$$

which is the probability that a sample from the sampling distribution will be in C. In general, C should be chosen to be as small as possible while still ensuring that both **1)** $F \cap H \subseteq C \subseteq G$ (so that informative samples are not discarded) and **2)** c can be computed. Ideally, we would select $C = F \cap H$, however in some cases c cannot be computed for this value of C. For example, in our later experiments we consider a problem where h and H are not known, but F is, and so we can compute c using $C = F$, but not $C = F \cap H$.

Let $k(\mathbf{X}_n) \coloneqq \sum_{i=1}^n \mathbf{1}_C(X_i)$ be the number of X_i that are in C. The US estimator is then defined as:

$$\mathrm{US}(\mathbf{X}_n) \coloneqq \frac{c}{k(\mathbf{X}_n)} \sum_{i=1}^n \frac{f(X_i)}{g(X_i)} h(X_i), \tag{2}$$

if $k(\mathbf{X}_n) > 0$, and $\mathrm{US}(\mathbf{X}_n) \coloneqq 0$ if $k(\mathbf{X}_n) = 0$. This is equivalent to applying importance sampling to the pruned data set, $\mathbf{X}'_n$, since then $g'(x) = g(x)/c$ for $x \in C$. Also, in (2) we sum over all n samples rather than just the $k(\mathbf{X}_n)$ samples in C because $f(X_i)h(X_i) = 0$ for all X_i not in C.

Although we analyze the US estimator as defined in (2), it can be generalized to use measure theoretic probability and to incorporate a control variate. In this more general setting, f and g are probability measures, f is absolutely continuous with respect to g, $t(X_i)$ denotes a real-valued sample-dependent control variate, and

$$\mathrm{US}(\mathbf{X}_n) \coloneqq \frac{g(C)}{k(\mathbf{X}_n)} \left(\sum_{i=1}^n \frac{df}{dg}(X_i)\Big(h(X_i) - t(X_i)\Big)\right) - \mathbf{E}_g[t(X)].$$

Theoretical Analysis of US

We begin with two simple theorems that elucidate the relationship between IS and US. The proofs of both theorems are straightforward, but deferred to the supplemental document. First, Theorem 1 shows that, when $C = G$, US degenerates to IS. One case where $C = G$ is when the support of the target distribution and evaluation function are both equal to the support of the sampling distribution, i.e., when $F = H = G$, and so $C = G$ necessarily.

Theorem 1. *If* $C = G$, *then* $\mathrm{US}(\mathbf{X}_n) = \mathrm{IS}(\mathbf{X}_n)$.

Theorem 2 shows that, if we replace c in the definition of US with an empirical estimate, $\hat{c}(\mathbf{X}_n) \coloneqq k(\mathbf{X}_n)/n$, then US and IS are equivalent. This provides some intuition for why US tends to outperform IS when $C \subset G$—IS is US, but using an empirical estimate of c (the probability that a sample falls within C), in place of its known value.

Theorem 2. *If we replace* c *with an empirical estimate,* $\hat{c}(\mathbf{X}_n) \coloneqq k(\mathbf{X}_n)/n$*, then* $\mathrm{US}(\mathbf{X}_n) = \mathrm{IS}(\mathbf{X}_n)$.

In Table 1 we summarize more theoretical results that clarify the differences between IS and US in several settings. The first setting (denoted by a † in Table 1) is the standard setting where we consider the ordinary expected value and variance of the two estimators. The second setting (denoted by a ‡ in Table 1) conditions on the event that at least one sample falls within C, that is, the event that $k(\mathbf{X}_n) > 0$. This is a reasonable setting to consider if one takes the view that no estimate should be returned if all of the samples are outside C. That is, if the pruned data set, $\mathbf{X}'_n$, is empty, then no estimate should be produced or considered (just as IS does not produce an estimate when $n = 0$—when there are no samples at all). Finally, the third setting (denoted by a $\star$ in Table 1) conditions on the event that $k(\mathbf{X}_n) = \kappa$—that a specific constant number of the n samples are in C.

Table 1 and the theorems that it references use additional symbols that we review here. Let $\rho \coloneqq \Pr(k(\mathbf{X}_n) > 0) = 1 - (1-c)^n$ be the probability that at least one of n samples is in C. Let $\mathrm{Var}_g(\cdot)$ denote the variance given that the random variables within the parenthesis are sampled from the distribution with PDF g. Let

$$v \coloneqq \mathrm{Var}_g\left(\frac{f(X)}{g(X)}h(X)\middle| X \in C\right)$$

be the conditional variance of the importance sampling estimate when using a single sample and given that the sample is in C. Let $B(n, c)$ denote the binomial distribution with parameters n and c and let $\mathbf{E}_{B(n,c)}$ denote the expected value given that $\kappa \sim B(n, c)$.

Although the proofs of the claims in Table 1 are some of the primary contributions of this work, we defer them to the supplemental document because they are straightforward (though lengthy) and do not provide further insights into the results. The primary result of Table 1 is that US is unbiased and often has lower variance in the key setting of interest: when at least one sample is in the support of the target distribution—when $k(\mathbf{X}_n) > 0$. We find this setting compelling because, when no samples are in F, little can be inferred about $\mathbf{E}_f[h(X)]$.

In this setting (denoted by ‡ in Table 1) US is an unbiased estimator, while IS is not (although the bias of IS does go to zero as $n \to \infty$).[1] To understand the source of this bias, consider the bias of IS given that $k(\mathbf{X}_n) = \kappa$—the $\star$ setting in Table 1. In this case, $\mathbf{E}_g[\mathrm{IS}(\mathbf{X}_n)] = \frac{\kappa}{cn}\theta$. Recall that IS uses an empirical estimate of c, i.e., $\hat{c} \approx \frac{\kappa}{n}$ (as discussed in Theorem 2). When this estimate is correct, terms in $\frac{\kappa}{cn}\theta$ cancel, making IS unbiased. Thus, the bias of IS when conditioning on the event that $k(\mathbf{X}_n) > 0$ stems from IS's use of an estimate of c.

Next we discuss the variance of the two estimators given that at least one sample falls within C, i.e., in the ‡ setting. First consider how the variances of IS and US change as $c \to 0$—that is, as the differences between the supports of the sampling and target distributions increases. Specifically, let $c_i \coloneqq \frac{1}{i}$ for $i \in \mathbb{N}_{>0}$. We then have that: $\mathrm{Var}(\mathrm{IS}(\mathbf{X}_n)|k(\mathbf{X}_n) > 0, c_i) \geq \frac{c_i v}{n\rho} = \frac{v}{n\rho i} \geq \frac{v}{ni}$, since $\rho \in (0, 1]$, and $\mathrm{Var}(\mathrm{US}(\mathbf{X}_n)|k(\mathbf{X}_n) > 0, c_i) =$

[1] If we do not condition on the event that $k(\mathbf{X}_n) > 0$, then US is a *biased* estimator of θ. This is because it is unclear how to define $\mathrm{US}(\mathbf{X}_n)$ when $k(\mathbf{X}_n) = 0$, and we chose (arbitrarily) to define it to be 0. However, the bias of $\mathrm{US}(\mathbf{X}_n)$ in this setting converges quickly to zero, since ρ (the probability that no samples fall within C) converges quickly to one as $n \to \infty$.

	$\mathbf{E}_g[\cdot]^\dagger$	$\mathbf{E}_g[\cdot]^\ddagger$	$\mathbf{E}_g[\cdot]^\star$	**Variance**†	**Variance**‡	**Strongly Consistent**
IS	θ (Property 1)	$\frac{1}{\rho}\theta$ (Theorem 6)	$\frac{\kappa}{cn}\theta$ (Theorem 5)	$\frac{1}{n}\left(cv+\theta^2\left(\frac{1}{c}-1\right)\right)$ (Theorem 11)	$v\frac{c}{n\rho}+\theta^2\frac{c\rho(n-1)+\rho-cn}{cn\rho^2}$ (Theorem 9)	Yes († and ‡)
US	$\rho\theta$ (Theorem 7)	θ (Theorem 4)	θ (Theorem 3)	$\rho c^2 v\mathbf{E}_{B(n,c)}[\kappa^{-1}\lvert\kappa>0]$ $+\theta^2\rho(1-\rho)$ (Theorem 10)	$c^2 v\mathbf{E}_{B(n,c)}[\kappa^{-1}\lvert\kappa>0]$ (Theorem 8)	Yes († and ‡)

Table 1: Theoretical properties of IS and US estimators. † = given no conditions. ‡ = conditioned on the event that $k(\mathbf{X}_n) > 0$—that at least one sample is in C. ⋆ = conditioned on the event that $k(\mathbf{X}_n) = \kappa$—that exactly κ of n samples are in C. All theorems require the assumption that $F \cap H \subseteq G$. The consistency results follow immediately from the fact that the biases and variances all converge to zero as $n \to \infty$ (Thomas and Brunskill 2016, Lemma 3).

$(v/i^2)\mathbf{E}_{B(n,c)}[1/\kappa|\kappa > 0] \leq v/i^2$, since $\mathbf{E}_{B(n,c)}[\kappa^{-1}|\kappa > 0] \leq 1$. Thus, as $i \to \infty$ (as $c \to 0$ logarithmically), and given some fixed n and v, the variance of US goes to zero much faster than the variance of IS. The variance of US (as a function of i) converges to zero linearly (or faster) with a rate of at most 1 while the variance of IS converges to zero sublinearly (at best, logarithmically).

Next note that the variance of US in this setting is independent of θ^2, but the variance of IS increases with θ^2 (see Property 3 in the supplemental document, applied to Theorem 9). To ameliorate this issue, a control variate, t, can be used to center the data so that $\theta \approx 0$. However, since θ is not known *a priori*, selecting $t = \theta$ is not practical. The term that scales with θ^2 in the variance of IS given that $k(\mathbf{X}_n) > 0$ therefore means that the variance of IS depends on the quality of the control variate—poor control variates can cause IS to have high variance. By contrast, the variance of US in this setting does not have a term that scales with θ^2, and so the quality of the control variate is less important.[2]

There is a rare case when IS can have a lower variance than US. First, we assume that the control variate is perfect so that $\theta = 0$ (which, as discussed before, is impractical) and consider the term that scales with v. From this term, it is clear that US will have lower variance than IS if:

$$c^2\mathbf{E}_{B(n,c)}[\kappa^{-1}|\kappa > 0] \leq \frac{c}{n\rho}. \tag{3}$$

Notice that this inequality depends only on n and c, which must both be known in order to implement US, and so we can test *a priori* whether US will have lower variance than IS. That is, if (3) holds, then US will have lower variance than IS, given that $k(\mathbf{X}_n) > 0$. However, if (3) does not hold, it does not mean that IS will have lower variance than US unless the perfect (typically unknown) control variate is used so that $\theta = 0$.

Application to Illustrative Example

Because neither method is always superior, here we consider the application of IS and US to the illustrative example to see when each method works best, and by how much. We consider the setting where $C = F$, but modify the example slightly. First, although the target distribution is always uniform, we allow for its support to be scaled. Specifically, we define the support of f to be $[0, F_{\text{max}}]$, where $F_{\text{max}} \in (0, 2]$. When F_{max} is small, it corresponds to significant differences in support, while large F_{max} correspond to small differences (when $F_{\text{max}} = 2$, $C = F = G$ and so the two estimators are equivalent). We also modify h to allow for various values of θ. Specifically, we define $h(x) = -1 + \theta$ if $x < F_{\text{max}}/2$ and $h(x) = 1 + \theta$ if $x \geq F_{\text{max}}/2$. Notice that, although we defined h in terms of θ, θ remains $\mathbf{E}_f[h(X)]$, and also that using this definition of h and $\theta = 0$ is an instance that is particularly favorable to IS.

For this example, it is straightforward to verify that $v = 4/F_{\text{max}}^2$ for any definition of θ, and $c = F_{\text{max}}/2$. Given these two values (and θ), we can compute the bias and variance of each estimator. The biases and variances of the two estimators for various settings are depicted in Figure 2. Notice that US is always competitive with IS, although the reverse is not true. Particularly, when F_{max} is small (so that c is small), or when θ is large, US can have orders of magnitude lower variance than IS. Also, as n increases, the two estimators become increasingly similar, since the empirical estimate of c used by IS becomes increasingly accurate, although US is still vastly superior to IS even when n is large if c is correspondingly small. This matches our theoretical analysis from the previous section: we expect US to perform better when c is small (by our convergence rate analysis) or when θ^2 is large (due to US's lesser dependence on the quality of the control variate), and we expect the two estimators to become increasingly similar as $n \to \infty$ (because $\hat{c}$ becomes increasingly similar to c).

Notice also that gains are not only obtained when c is so small relative to n that no samples are expected to fall within C (a relatively uninteresting setting). For example, the right-most plot in Figure 2 shows that with $F_{\text{max}} = 0.5$, where $\Pr(k(\mathbf{X}_n) > 0) = \rho = 1 - \frac{1}{2^{50}} \approx 1$, the MSE of US is approximately 0.086, while the MSE of IS is approximately 6.08—US is has roughly $1/70$ the MSE of IS ($1/8$ the RMSE).

Perhaps surprisingly, there are cases where IS has lower variance than US (even when both are unbiased, since $\theta = 0$). For example, consider the plot with $\theta = 0$ and $n = 10$, and the position on the horizontal axis that corresponds to $F_{\text{max}} = 1.0$. This is one case where IS is marginally better than US (it has lower variance in both settings, and neither estimator is biased). Intuitively, the IS estimator includes the points outside the support of F, although they have associated values, $h(X_i) = 0$, which pulls the importance sam-

[2] The quality of the control variate can still impact the variance of estimates though, since it can change v.

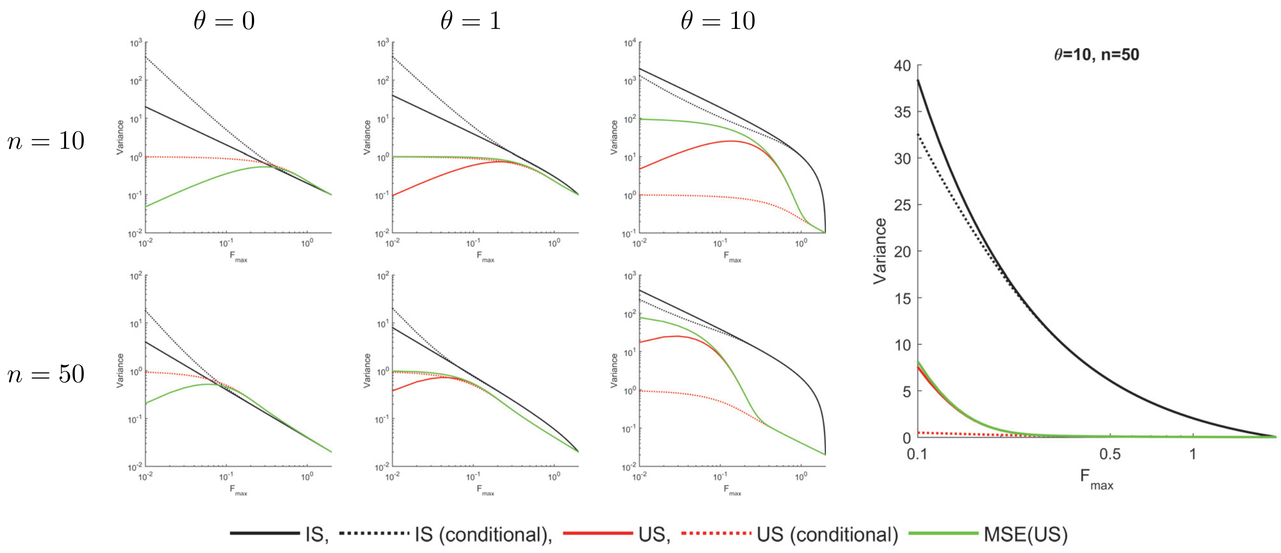

Figure 2: The variances of IS and US across various settings of n and θ (denoted along the left and top). At a glance, notice that the red and green curves (US) tend to be below the black curves (IS), particularly when considering the logarithmic scale of the vertical axes. The dotted lines show the variance conditioned on the event that $k(\mathbf{X}_n) > 0$. The green line shows the mean squared error of the US estimator (without any conditions), which shows that the variance reduction of US is not completely offset by increased bias (compare the solid black and green curves). When $\theta = 0$ the green line obscures the solid red line. The plot on the right shows a zoomed-in view of the $\theta = 10, n = 50$ plot without the logarithmic vertical axis.

pling estimate towards zero. In this case, when $\theta = 0$, this extra pull towards zero happens to be beneficial. However, to remain unbiased given the pull towards zero, IS also increases the magnitudes of the weights associated with points in F, which incurs additional variance. When F_{max} is small enough, this additional variance outweighs the variance reduction that results from the extra pull towards zero, and so US is again superior. This intuition is supported by the fact that in Figure 2 IS does not outperform US for small F_{max} or $\theta \geq 1$, since then a pull towards zero is detrimental.

Finally, we consider the use of IS and US to create high-confidence upper and lower bounds on θ using a concentration inequality (Massart 2007) like Hoeffding's inequality (Hoeffding 1963). If b denotes the range of the function $f(x)h(x)/g(x)$, for $x \in G$, then using Hoeffding's inequality, we have that $\text{IS}(\mathbf{X}_n) - b\sqrt{\ln(1/\delta)/(2n)}$ is a $1 - \delta$ confidence lower bound on θ. Similarly, we can use US with Hoeffding's inequality to create a $1 - \delta$ confidence lower bound: $\text{US}(\mathbf{X}_n) - cb\sqrt{\ln(1/\delta)/(2k(\mathbf{X}_n))}$, since the range of the $k(\mathbf{X}_n)$ i.i.d. random variables averaged by $\text{US}(\mathbf{X}_n)$ is cb. Notice that, if $k(\mathbf{X}_n) = 0$, then this second estimator is undefined (one might define the lower bound to be a known lower bound on θ in this setting). Although we expect that $k(\mathbf{X}_n) \approx cn$, the resulting c in the denominator of the US-based bound is within the square root, while the c in the numerator is not, and so the bound constructed using US should tend to be tighter when c is small.

Application to Diabetes Treatment

We applied US and IS to the problem of predicting the effectiveness of altering the treatment policy for a particular person with type 1 diabetes. That is, we would like to use prior data from when the individual was treated with one treatment policy to estimate how well a related policy would work. The treatment policy is parameterized by two numbers, CR and CF, and dictates how much insulin a person should inject prior to eating a meal in order to keep his or her blood glucose close to optimum levels. CR and CF are typically specified by a diabetologist and tweaked during follow-up visits every 3–6 months. If follow-up visits are not an option, recent research has suggested using reinforcement learning algorithms to tune CR and CF (Bastani 2014).

Here we focus on a sub-problem of improving CR and CF—using data collected from an initial range of admissible values of CR and CF to predict how well a new range of values for CR and CF would perform. When collecting data, CR and CF are drawn uniformly from an initial admissible range, and then used for one day (which we view as one episode of a Markov decision process). The performance during each day is measured using an objective function similar to the reward function proposed by Bastani (2014), which measures the deviation of blood glucose from optimum levels, with larger penalties for low blood glucose levels. We refer to the measure of how good the outcome was from one day as the *return* associated with that day, with larger values being better. Using approximately 30 days of data, our goal is to estimate the expected return if a different distribution of CR and CF were to be used.

We consider a specific *in silico* person—a person simulated using a metabolic simulator. We used the subject "Adult#003" in the Type 1 Diabetes Metabolic Simulator (T1DMS) (Dalla Man et al. 2014)—a simulator that has

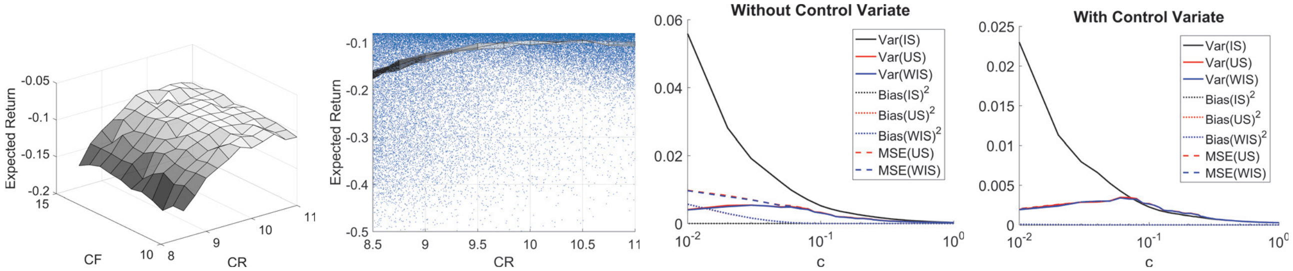

Figure 3: The first and second plots show an estimate of the expected return for various CR and CF, from two different angles (the second is a side-view of the first). The second plot also includes blue points depicting the Monte Carlo returns observed from using different values of CR and CF for a day—notice the high variance. The two plots on the right depict the bias, variance, and MSE of IS, US, and WIS (without any conditioning) for various values of c and both without (third plot) and with (fourth plot) a control variate. The curves for US are largely obscured by the corresponding curves for WIS. Notice that the variance of IS approaches 0.06, which is enormous given that the difference between the best and worst CR and CF pairs possible under the sampling policy is approximately 0.06.

been approved by the US Food and Drug Administration as a substitute for animal trials in pre-clinical testing of treatment policies for type 1 diabetes. During each day, the subject is given three or four meals of randomized sizes at randomized times, similar to the experimental setup proposed by Bastani (2014). As a result of this randomness, and the stochastic nature of the T1DMS model, applying the same values of CR and CF can produce different returns if used for multiple days. After analyzing the performance of many CR and CF pairs, we selected an initial range that results in good performance: CR $\in [8.5, 11]$ and CF $\in [10, 15]$. Using a large number of samples, we computed a Monte Carlo estimate of the expected return if different CR and CF values are used for a single day—this estimate is depicted in Figure 3.

As described by Bastani (2014), when the value of CR is set appropriately, performance is robust to changes in CF. We therefore focus on possible changes to CR. Specifically, we consider new treatment policies where CF remains sampled from the uniform distribution over $[10, 15]$, but where CR is sampled from the truncated normal distribution over $[\text{CR}_{\text{min}}, 11]$, with mean 11 and standard deviation $11 - \text{CR}_{\text{min}}$. This distribution places the largest probability densities at the upper end of the range of CR, which favors better policies. As CR_{min} increases towards 11, the support of the sampling distribution and target distribution become increasingly different ($c = (11 - \text{CR}_{\text{min}})/2.5$) and the expected return increases.

For each value of CR_{min} (each of which corresponds to a value of c), we performed $2{,}433$ trials, each of which involved generating the returns from 30 days, where the values of CR and CF used for each day were sampled uniformly from CR $\in [8.5, 11]$ and CF $\in [10, 15]$, and then using IS, US, and *weighted importance sampling* (WIS) to estimate the expected return if CR and CF were sampled from the target distribution (the truncated Gaussian parameterized by CR_{min}). Figure 3 displays the bias, variance and *mean squared error* (MSE) of these $2{,}433$ estimates, using an estimate of ground truth computed using Monte Carlo sampling. Figure 3 also shows the impact of providing a constant control variate to all the estimators: the chosen control variate was the expected return under the sampling distribution.

Notice that we see the same trend as in the illustrative example—for small c (the best treatment policies, which have small ranges of CR), US significantly outperforms IS. Furthermore, when a decent control variate is not used, the benefits of US are increased, even when controlling for the resulting bias by measuring the mean squared error. We also computed the biases and variances given that $k(\mathbf{X}_n) > 0$, and observed similar results (not shown), which favored US slightly more. Notice that WIS and US perform very similarly. Indeed, if the sampling and target distributions are both uniform, it is straightforward to verify that WIS and US are equivalent. In other experiments (not shown) we found that WIS yields lower variance than US when the target distribution is modified to be even less like the uniform distribution.

However, it is often important to be able to produce confidence intervals around estimates (especially when data is limited), and since WIS is biased, it cannot be used with standard concentration inequalities. We used Hoeffding's inequality to compute a 90% confidence interval around the estimates produced by IS and US (without control variates and with $\text{CR}_{\text{min}} = 10.375$, so that $c = 1/4$) using various numbers of samples (days of data). The mean confidence intervals are depicted in Figure 4, which also shows a Monte Carlo estimate of θ, as well as deterministic domain-specific upper and lower bounds on $h(X)$ (denoted by "h range" in the legend). If $k(\mathbf{X}_n) = 0$, then US is not defined, and so the confidence intervals shown for US are averaged only over the instances where $k(\mathbf{X}_n) > 0$. To show how often US returns a solution, Figure 4 also shows ρ—the probability that US will produce a confidence bound—using the right vertical axis for scale.

US produces a much tighter confidence interval than IS in all cases. Furthermore, the setting where US often does not return a bound corresponds to the setting where IS produces a confidence interval that is outside the deterministic bound on $h(X)$—a trivial confidence interval. In additional experiments (not shown) we defined the bounds to be truncated to always be within the deterministic bounds on $h(X)$ and define the bound produced using US to be conservative

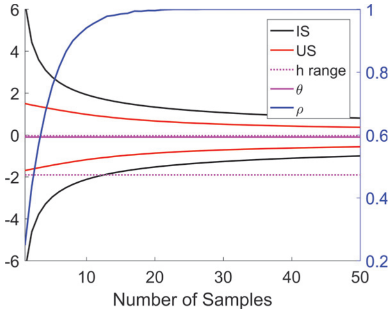

Figure 4: Confidence bounds using IS and US.

(equal to the deterministic bounds) when $k(\mathbf{X}_n) = 0$. In this experiment we saw similar results—the confidence intervals produced using US were much tighter than those using IS.

Should One Use US or WIS in Practice?

The results presented in the previous section might raise the question: when should one use US rather than WIS? Previously we hinted at the problem with WIS: it is a biased estimator. Here we discuss why this theoretical property has important practical ramifications that rule out the use of WIS (but not US) for many high-risk problems.

First we list the troublesome theoretical properties of the WIS estimator, which are discussed in the work of Bastani (2015). When there is only a single sample, i.e., when $n = 1$, WIS is an unbiased estimator of $\mathbf{E}_g[h(X)]$. As n increases, the expected value of the WIS estimator shifts towards the target value, $\theta = \mathbf{E}_f[h(X)]$. If the samples that are likely under g are extremely unlikely under f, then the shift of the expected value of the WIS estimator from $\mathbf{E}_g[h(X)]$ to $\mathbf{E}_f[h(X)]$ can be exceedingly slow.

Consider what this would mean for our diabetes experiment. Here the behavior policy (sampling distribution) is a relatively decent policy that we might be considering changing. The evaluation policy (target distribution) might be a new treatment policy that is both dangerously worse than the behavior policy and quite different from the behavior policy. To determine whether the evaluation policy should be deployed, we might rely on high-confidence guarantees, as has been suggested for similar problems (Thomas et al. 2015a). That is, we might use Hoeffding's inequality to construct a high-confidence lower-bound on the expected value of the WIS estimator, and then require this bound to be not far below the performance of the behavior policy.

Because the behavior and evaluation policies are quite different, the WIS estimator will produce relatively low-variance estimates centered near the performance of the reasonable behavior policy, rather than estimates centered near the dangerously poor performance of the evaluation policy. This means that the lower-bound that we compute will be a lower bound on the performance of the decent behavior policy, rather the true poor performance of the evaluation policy. Moreover, if one uses Student's t-test or a bootstrap method to construct the confidence interval, as has been suggested when using WIS (Thomas et al. 2015b), we might obtain a very-tight confidence interval around the performance of the behavior policy. This exemplifies the problem with using WIS for high-risk problems: the bias of the WIS estimator can cause us to often erroneously conclude that dangerous policies are safe to deploy.

Conclusion and Future Work

We have presented a simple new variant of importance sampling, US. Our analytical and empirical results suggest that US can significantly outperform ordinary importance sampling when the supports of the sampling and target distributions differ. We also provide an inequality that can be evaluated prior to observing any data, and which, if satisfied, guarantees that US will have lower variance than ordinary importance sampling. Unlike some other importance sampling estimators that have been developed to reduce variance (like WIS), US is unbiased given mild conditions that still permit the easy computation of confidence intervals.

References

Bastani, M. 2014. Model-free intelligent diabetes management using machine learning. Master's thesis, Department of Computing Science, University of Alberta.

Dalla Man, C.; Micheletto, F.; Lv, D.; Breton, M.; Kovatchev, B.; and Cobelli, C. 2014. The UVA/Padova type 1 diabetes simulator new features. *Journal of Diabetes Science and Technology* 8(1):26–34.

Dudík, M.; Langford, J.; and Li, L. 2011. Doubly robust policy evaluation and learning. In *ICML*, 1097–1104.

Hammersley, J. M., and Handscomb, D. C. 1964. Monte Carlo methods, Methuen & Co. *Ltd., London* 40.

Hammersley, J. M. 1960. Monte Carlo methods for solving multivariable problems. *Annals of the New York Academy of Sciences* 86(3):844–874.

Hoeffding, W. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58(301):13–30.

Jiang, N., and Li, L. 2016. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*.

Kahn, H. 1955. Use of different Monte Carlo sampling techniques. Technical Report P-766, The RAND Corporation.

Massart, P. 2007. *Concentration Inequalities and Model Selection*. Springer.

Rubinstein, R. 1981. *Simulation and the Monte Carlo method*. New York: Wiley.

Thomas, P. S., and Brunskill, E. 2016. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*.

Thomas, P. S.; Theocharous, G.; ; and Ghavamzadeh, M. 2015a. High confidence off-policy evaluation. In *AAAI*.

Thomas, P. S.; Theocharous, G.; ; and Ghavamzadeh, M. 2015b. High confidence policy improvement. In *International Conference on Machine Learning*.

Thomas, P. S. 2015. *Safe Reinforcement Learning*. Ph.D. Dissertation, University of Massachusetts Amherst.